
ELSEVIER

Journal of Computational and Applied Mathematics 91 (1998) 107–122

JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS

Difference equation models for estimating athletic records

G.R. Dargahi-Noubary *, Yixun Shi

Department of Mathematics and Computer Science, Bloomsburg University of Pennsylvania, Bloomsburg, PA 17815, USA

Received 2 September 1997; received in revised form 5 February 1998

Abstract

This article presents an analysis of the models that are frequently used for estimating the future records for atheletic events. It considers an innovative approach for modeling based on difference equations whose complementary solution yields one of these models. It does this by establishing a relationship between the models and certain difference equations. Moreover, it shows that each of those models can be converted to the exponential model by applying an appropriate substitution. Based on the derived relations it proposes a numerical procedure for estimating the model parameters and discusses its advantages over an existing procedure. This procedure reduces the computational cost and avoids the classical difficulties of non-linear optimization procedures such as the Newton's method. Finally, it tests the proposed numerical procedure using simulated data and applies the procedure to the fastest times for 400 m race.

Keywords: Difference equation; Athletic records; Exponential model; Numerical procedure

1. Introduction

Apart from intrinsic interest, there are several medical and physiological reasons why one would like to know how fast a human being could, for example, run a medium distance such as 400 m. The world record for the 400 m, 43.29 s was made on 17 August 1988 by Butch Reynolds. To date, this record has not been bettered but it will happen one day. While there is a general agreement among the Physiologists and Physical educators about existence of an upper limit for such a speed (or a lower limit for the time needed to run this distance) the limit is not known at the present time. Due to a great interest in this question, apart from physiological research there have been some attempts to estimate (predict) the limits through mathematical and statistical modeling. Notable efforts include work of Chatterjee and Chatterjee [2], Tryfos and Blackmore [9], and Smith [8]. A general

* E-mail: noubary@planetx.bloomu.edu

0377-0427/98/$19.00
PII S0377-0427(98)00032-6

approach employed in these studies has been based on predictive mathematical models via examining relationships between the past and present records.

Smith [8] has applied a so-called exponential-decay model to the records for mile and marathon races. Unfortunately, when applying the model to estimate the limiting record (best attainable time) the standard errors were so large that the estimates were meaningless. Chatterjee and Chatterjee [2] have applied a similar model to the winning times in the men's 100, 200, 400 and 800 m runs in the Olympic Games. They have obtained two sets of estimates for the parameters of interest and for the ultimate winning record using non-linear model fitting algorithm and jack-knife procedure.

More recently, Dargahi-Noubary [3] has shown that the use of an alternative model involving logarithms of the annual records could improve the analysis of the winning times and has used the results for short-time prediction of the men's 400 and 800 m run.

Most of the successful models proposed so far are based on the following form:

$$Y(t) = Z(t, \boldsymbol{\theta}) + x(t), \tag{1}$$

where $Y(t)$ represents records, $Z(t, \boldsymbol{\theta})$ is a deterministic function (smooth part) representing the trend and $x(t)$ is the random component (rough part) representing the fluctuations. Blest [1] has considered some frequently used models for $Z(t, \boldsymbol{\theta})$ and has examined their performance using records at the time of Olympics. These models, together with the antisymmetric exponential model introduced in Blest [1] are listed in Table 1.

The present article suggests consideration of an innovative approach for modeling and estimating the records based on difference equations whose complementary solution yields models listed in the Table 1. Since for each model we have one difference equation and vice versa, it is possible to carry the analysis using difference equations alone or in combination with the above models. Since all the above models present smooth functions, such analysis could also be performed based on differential equations measured at the integer time values.

2. Derivation of difference equations

In this section we derive difference equations for each model listed in Table 1. Let $y(n)$ represent, for example, the fastest time for 400 m run at time n or at year n. The linear model (straight line) $\theta_1 - \theta_2 t$ satisfies the following difference equation:

$$y(n) = y(n-1) - \theta_2. \tag{2}$$

Note that for this case $y(n) - y(n-1) = -\theta_2$ which represents a constant annual improvement. This implies that $y(n) = y(1) - (n-1)\theta_2$, resulting in negative values when n gets large. To avoid this, we may consider the following assumptions:

(I) The annual improvement, denoted by $y(n-1) - y(n)$, should in general decrease and eventually vanish.

(II) If $\lim_{n\to\infty} y(n) = y_*$, then in general $\{y(n)\}$ decreases monotonically to y_* and hence $y(n) - y_*$ decreases monotonically to zero.

Based on these assumptions, we will consider a few difference equations corresponding to the models listed in Table 1.

Table 1
Frequently used models reproduced from Blest [1]

Model	Comment
$Z(t,\theta)=\theta_1+\theta_2 t$	Straight line
$Z(t,\theta)=\theta_1+\theta_2\exp(-\theta_3 t)$	Exponential
$Z(t,\theta)=\delta-\alpha\{1-\exp(-\beta t)\}^{\gamma}$	Extended Chapman–Richards; see Ratkowsky [7]
$Z(t,\theta)=\delta+\alpha\exp\{-\beta(t-\gamma)\}$ for $t\geqslant\gamma$	Antisymmetric
$Z(t,\theta)=\delta+\alpha[2-\exp\{\beta(t-\gamma)\}]$ for $t<\gamma$	Exponential
$Z(t,\theta)=\delta-\alpha\{1+\exp(\beta-\gamma t)\}^{-1}$	Logistic
$Z(t,\theta)=\delta-\alpha\exp\{-\exp(\beta-\gamma t)\}$	Four-parameter Gompertz; see Ratkowsky [7]
$Z(t,\theta)=\delta+\alpha\exp\{-\exp[\gamma(t-\varepsilon)]\}$	Reparameterization of above; see Ratkowsky [7]

First, we derive a difference equation for the exponential model. Suppose that

$$y(n-1)-y(n)=r(y(n-2)-y(n-1)) \quad \forall n, \tag{3}$$

where $r\in(0,1)$ is a constant. If we rewrite (3) as

$$y(n)-(1+r)y(n-1)+ry(n-2)=0, \tag{4}$$

then the characteristic equation of (4), namely

$$p^2-(1+r)p+r=0 \tag{5}$$

has two solutions $p=1$ and $p=r$. Hence, the general solution of (3) is of the form

$$y(n)=a+br^n,$$

where a and b are constant coefficients. Since $\{y(n)\}$ is in general decreasing, it is reasonable to assume that b is positive. Thus, if we let $a=\theta_1, b=\theta_2>0$, and $r=\exp(-\theta_3)$, then we arrive at exponential model

$$y(n)=\theta_1+\theta_2\exp(-\theta_3 n) \tag{6}$$

that is, the exponential model can be derived from the difference equation (3).

Let us now consider a difference equation corresponding to the Extended Chapman–Richards model. Suppose $\alpha>0$ is a fixed number and let $\delta=y_*+\alpha$. Based on assumption (II) of this section, we may assume that when n is sufficiently large $((\delta-y(n))/\alpha)^{1/\gamma}$ is increasing for any $\gamma>0$. Furthermore, if we assume that

$$((\delta-y(n))/\alpha)^{1/\gamma}-((\delta-y(n-1))/\alpha)^{1/\gamma}=r[((\delta-y(n-1))/\alpha)^{1/\gamma}-((\delta-y(n-2))/\alpha)^{1/\gamma}], \tag{7}$$

where $r\in(0,1)$ is constant, then we obtain the difference equation we are seeking. To solve this equation we let $w(n)=((\delta-y(n))/\alpha)^{1/\gamma}$. Eq. (7) can be written as

$$w(n-1)-w(n)=r(w(n-2)-w(n-1)) \tag{8}$$

which takes the same form as Eq. (3). Thus, noting that $w(n)$ is increasing, the solution of (8) is given by

$$w(n) = \theta_1 - \theta_2 \exp(-\theta_3 n). \tag{9}$$

Here θ_1 and $\theta_2 > 0$ are coefficients and $\theta_3 = -\ln r > 0$. If we choose $\theta_1 = \theta_2 = 1$ and denote $\beta = \theta_3$, then from (9) we have $w(n) = 1 - \exp(-\beta n)$, which implies that

$$y(n) = \delta - \alpha[1 - \exp(-\beta n)]^{\gamma}. \tag{10}$$

This is the Extended Chapman–Richards model.

Turning to Antisymmetric Exponential model, we note that it is equivalent to the piecewise exponential model

$$y(n) = \delta + \alpha \exp(\beta\gamma) \exp(-\beta n) \quad \text{when } n \geqslant \gamma$$

and

$$y(n) = (\delta + 2\alpha) - \alpha \exp(-\beta\gamma) \exp(\beta n) \quad \text{when } n < \gamma.$$

Hence the corresponding difference equations can easily be derived from the piecewise difference equations

$$y(n-1) - y(n) = r(y(n-2) - y(n-1))$$

with

$$r = \exp(-\beta) \quad \text{when } n \geqslant \gamma$$

and

$$r = \exp(\beta) \quad \text{when } n < \gamma.$$

For the Logistic model, we once more use the notation $\delta = y_* + \alpha$ where $\alpha > 0$ is a constant. Suppose that

$$y(n-1) - y(n) = r(y(n-2) - y(n-1))$$

but now r is a function defined as

$$r = a \frac{y(n) - \delta}{y(n-1) - \delta} \cdot \frac{y(n-1) - \delta}{y(n-2) - \delta} \tag{11}$$

with $a \in (0, 1)$ being a constant. That is, we consider the difference equation

$$y(n-1) - y(n) = a \frac{y(n) - \delta}{y(n-1) - \delta} \cdot \frac{y(n-1) - \delta}{y(n-2) - \delta} (y(n-2) - y(n-1)) \tag{12}$$

In order to solve (12) we use a substitution $w(n) = (y(n) - \delta)^{-1}$. Notice now that $w(n)$ is increasing. With this substitution, Eq. (12) becomes

$$w(n-1) - w(n) = a(w(n-2) - w(n-1))$$

which has the solution $w(n) = \theta_1 - \theta_2 \exp(-\theta_3 n)$ with $\theta_2 > 0$ and $\theta_3 = -\ln a > 0$. Also, since $w(n) \to -1/\alpha < 0$ we must have that $\theta_1 < 0$. This implies that

$$y(n) = \delta + [\theta_1 - \theta_2 \exp(-\theta_3 n)]^{-1}$$

with $\theta_1 < 0$. Let $\theta_1 = -1/\alpha$, $\beta = \ln \theta_2/(-\theta_1)$, and $\gamma = \theta_3$, we then have

$$y(n) = \delta - \alpha\{1 + \exp(\beta - \gamma n)\}^{-1} \tag{13}$$

which yields the formula of the Logistic model.

It is interesting to note that the classical Logistic model (13) is derived indirectly from the Discrete Logistic Equation

$$y(n+1) - y(n) = r^* y(n)(1 - y(n)/k),$$

where r^* and k are constants. The continuous version of the discrete Logistic equation is a first-order Bernoulli differential equation

$$\frac{dy}{dt} = ay - by^2$$

with $a = r^*$ and $b = r^*/k$. The general solution of this differential equation takes the form

$$y = (b/a + c \exp(-at))^{-1},$$

where c is a constant. Let $\alpha = -a/b, \gamma = a$, and $\beta = \ln((a/b)c)$, we then can write this solution as

$$y = -\alpha\{1 + \exp(\beta - \gamma t)\}^{-1}.$$

By considering the discrete version of the above solution with a shift δ, we obtain the formula for the Logistic model, namely

$$y(n) = \delta - \alpha\{1 + \exp(\beta - \gamma n)\}^{-1}.$$

In this paper, however, we showed that the Logistic model can be derived directly from the difference equation (12), which is based on the ratio of improvements.

Similarly, by considering the difference equation

$$w(n-1) - w(n) = a(w(n-2) - w(n-1))$$

with $a \in (0,1)$ and $w(n) = \ln(\delta - y(n))$ which is again an increasing sequence, we can using appropriate notations, get

$$w(n) = \ln \alpha - \exp(\beta) \exp(-\gamma n).$$

This leads to the 4-parameter Gompertz model

$$y(n) = \delta - \alpha \exp\{-\exp(\beta - \gamma n)\}. \tag{14}$$

Finally, if we denote $\delta = y_*$ and let $w(n) = \ln(y(n) - \delta)$, then $w(n)$ is decreasing and approaches $-\infty$. Again consider the difference equation

$$w(n-1) - w(n) = r(w(n-2) - w(n-1)),$$

Table 2
Improvement patterns

Model	Imp(n)
Exponential	$y(n) - y(n-1)$
Extended Chapman-Richards	$((\delta - y(n))/\alpha)^{1/\gamma} - ((\delta - y(n-1))/\alpha)^{1/\gamma}$
Antisymmetric Exponential	$y(n) - y(n-1)$
Logistic	$(y(n) - \delta)^{-1} - (y(n-1) - \delta)^{-1}$
Four-parameter Gompertz	$\ln(\delta - y(n)) - \ln(\delta - y(n-1))$
Reparameterization of Gompertz model	$\ln(y(n) - \delta) - \ln(y(n-1) - \delta)$

but now assume that $r > 1$. The general solution is now

$$w(n) = a + br^n$$

with $b < 0$. In this case, by denoting $a = \ln \alpha$, $b = -\exp(-\gamma\varepsilon)$, and $\gamma = \ln r$, we obtain the formula of the Reparameterization of 4-parameter Gompertz model

$$y(n) = \delta + \alpha \exp\{-\exp[\gamma(n - \varepsilon)]\}. \tag{15}$$

With that, we have shown that all the models listed in Table 1 can be derived from an appropriate difference equation. Furthermore, we have seen that they all can be converted into an exponential model through a proper substitution (by using $w(n) = \exp(y(n))$, the linear model $y(n) = \theta_1 + \theta_2 n$ can be easily converted into an exponential model).

The significance of the relation between difference equations and mathematical models is that this relation may help us in determining which model to use. For example, if the time series $\{y(n)\}$ approximately satisfies

$$\frac{y(n) - y(n-1)}{y(n-1) - y(n-2)} = \frac{y(n-1) - y(n-2)}{y(n-2) - y(n-3)} \quad \forall n > 3, \tag{16}$$

then we know that an exponential model is appropriate. As another example, if (16) holds for $w(n) = (y(n) - \delta)^{-1}$, that is if

$$\frac{w(n) - w(n-1)}{w(n-1) - w(n-2)} = \frac{w(n-1) - w(n-2)}{w(n-2) - w(n-3)} \quad \forall n > 3, \tag{17}$$

where δ is a constant, then the Logistic model is a proper model to use.

We also like to mention that all the difference equations we used in the derivation are based on an universal idea that tries to model the ratio of improvements. Different improvements patterns considered for different models are listed in Table 2. Here "Imp(n)" stands for the improvement considered at the nth term, or year.

In the following section, we shall present a numerical approach for estimating the parameters in some of the above models.

3. Numerical approach

All the models discussed in previous sections have the general form

$$y(n) = Z(n,\boldsymbol{\theta}) + x(n), \quad n = 1,2,\ldots,N \tag{18}$$

where $\{y(n)\}$ is the recorded time series, $x(n)$'s have mean value zero and stand for deviation from the expectation, and $Z(n,\boldsymbol{\theta})$ is the deterministic component. For exponential model, $Z(n,\boldsymbol{\theta}) = \theta_1 + \theta_2 \exp(-\theta_3 n)$ where θ_1, θ_2, and θ_3 are parameters to be estimated. For Logistic model, $Z(n,\boldsymbol{\theta}) = \delta - \alpha\{1 + \exp(\beta - \gamma n)\}^{-1}$ with δ, α, β, and γ being parameters. To estimate these parameters, we can use least-squares approach, that is

$$\min_{\boldsymbol{\theta}} \sum_{n=1}^{N} [y(n) - Z(n,\boldsymbol{\theta})]^2 \tag{19}$$

and employ the Newton's method to solve for the solution $\boldsymbol{\theta}$ from the following nonlinear system

$$\frac{\partial}{\partial \boldsymbol{\theta}} \sum_{n=1}^{N} [y(n) - Z(n,\boldsymbol{\theta})]^2 = 0. \tag{20}$$

Now, it is well known that Newton's method may fail if the initial guess for $\boldsymbol{\theta}$ is poor. Furthermore, this method is computationally very costly since the Jacobi matrix needs to be computed and a linear system has to be solved in each iteration. Due to the structure of (20), the computing of the Jacobi matrix alone will require a significant amount of computation for large N. To avoid this, in this section, we present a new approach for estimating parameters of some models discussed in previous sections. This approach may not only reduce the amount of computation cost but also save us the trouble of making the initial guess.

First, we consider the exponential model, that is,

$$y(n) = \theta_1 + \theta_2 \exp(-\theta_3 n) + x(n), \quad n = 1,2,\ldots,N. \tag{21}$$

Obviously, when n is large, the term $\theta_2 \exp(-\theta_3 n)$ is close to zero, and hence, we have

$$y(n) \doteq \theta_1 + x(n). \tag{22}$$

Considering this we may compute an estimate $\hat{\theta}_1$ of θ_1 as

$$\hat{\theta}_1 = \frac{\sum_{n=N-k+1}^{N} y(n)}{k}, \tag{23}$$

where k is an integer. Note that k has to be quite small for (22) to be valid, but not be too small for $\hat{\theta}_1$ to be a good estimate. One possible question here is: can we simply estimate $\hat{\theta}_1$ at $n = N$, that is to take $\hat{\theta}_1 = y(N)$? This in general is not quite appropriate due to the error part $x(N)$. But (23) offers a better estimation because now the error term is $\frac{\sum_{n=N-k+1}^{N} x(n)}{k}$ so that the error would be "averaged out" or reduced provided that k is not too small. In our numerical experiment presented later in this section, we chose k to be the integer part of one-tenth of the data size, that is, $k = [N/10]$.

Replacing $\hat{\theta}_1$ for θ_1 we then have

$$y(n) - \hat{\theta}_1 \doteq \theta_2 \exp(-\theta_3 n) + x(n) \quad n = 1, 2, \ldots, N. \tag{24}$$

Using the same idea and noting that

$$\frac{\sum_{n=1}^{N-1} \left(y(n) - \hat{\theta}_1\right)}{N-1} \doteq \frac{\theta_2 \sum_{n=1}^{N-1} \exp(-\theta_3 n)}{N-1}$$

and

$$\frac{\sum_{n=2}^{N} \left(y(n) - \hat{\theta}_1\right)}{N-1} \doteq \frac{\theta_2 \sum_{n=2}^{N} \exp(-\theta_3 n)}{N-1}$$

we may estimate $\hat{\theta}_3$ of θ_3 as

$$\hat{\theta}_3 = \ln \left[\frac{\sum_{n=1}^{N-1} \left(y(n) - \hat{\theta}_1\right)}{\sum_{n=2}^{N} \left(y(n) - \hat{\theta}_1\right)} \right]. \tag{25}$$

Finally, an estimate $\hat{\theta}_2$ of θ_2 can be obtained as

$$\hat{\theta}_2 = \left[\sum_{n=1}^{N} \left(y(n) - \hat{\theta}_1\right) \right] \exp\left(\hat{\theta}_3\right) \cdot \frac{1 - \exp\left(-\hat{\theta}_3\right)}{1 - \exp\left(-N\hat{\theta}_3\right)}$$

$$\doteq \left[\sum_{n=1}^{N} \left(y(n) - \hat{\theta}_1\right) \right] \left(\exp\left(\hat{\theta}_3\right) - 1\right). \tag{26}$$

It should be mentioned that the cost of computing $\hat{\theta}_1$, $\hat{\theta}_2$, and $\hat{\theta}_3$ is almost none compared to the Newton's method. Moreover, when using this method no initial guess is needed. When the error terms $x(n)$ are relatively small, $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}_3$ can furnish very accurate estimates for θ_1, θ_2, θ_3. In general, when error terms are not negligible, we can combine this approach with Newton's method. That is, use $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}_3$ as initial guess for Newton's method. Since the computational cost for $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}_3$ is low and that a good initial guess will lead to a fast convergence of Newton's method, this combination can reduce the total computational cost significantly. In practice, there are cases when even $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}_3$ as initial guess are not good enough for Newton's method to converge. If that happens, we can apply a smoothness procedure to the data set before executing this combination. Details of this is presented later with our numerical experiments.

Similar ideas can be applied to some other models listed in Table 1, based on the relation between these models and exponential model. For example, let us consider the Logistic model

$$Z(n, \boldsymbol{\theta}) = \delta - \alpha\{1 + \exp(\beta - \gamma n)\}^{-1}. \tag{27}$$

As discussed in Section 2, the substitution $w(n) = (y(n) - \delta)^{-1}$ converts this to an exponential model

$$w(n) = \theta_1 - \theta_2 \exp(-\theta_3 n). \tag{28}$$

Thus, for each estimate $\hat{\delta}$ of δ, the above approach can be used to get a set of estimates $\hat{\theta}_1(\hat{\delta})$, $\hat{\theta}_2(\hat{\delta})$, and $\hat{\theta}_3(\hat{\delta})$ which further yields the estimates

$$\hat{\alpha}(\hat{\delta}) = -1/\hat{\theta}_1(\hat{\delta}), \quad \hat{\beta}(\hat{\delta}) = \ln\left[\frac{\hat{\theta}_2(\hat{\delta})}{-\hat{\theta}_1(\hat{\delta})}\right], \quad \hat{\gamma}(\hat{\delta}) = \hat{\theta}_3(\hat{\delta}).$$

Now, if we denote $\boldsymbol{\theta}(\hat{\delta}) = (\hat{\delta}, \hat{\alpha}(\hat{\delta}), \hat{\beta}(\hat{\delta}), \hat{\gamma}(\hat{\delta}))$, then we can define a function

$$F(\hat{\delta}) = \sum_{n=1}^{N} \left[y(n) - Z(n, \boldsymbol{\theta}(\hat{\delta}))\right]^2 \tag{29}$$

and solve the reduced one-dimensional minimization problem

$$\min_{\hat{\delta}} F(\hat{\delta}) \tag{30}$$

to produce a fine estimate for parameters. Again, the computational cost of this approach is relatively low, and in general the estimates produced by this approach can serve as an initial guess for Newton's method.

Similarly, when we estimate parameters for Antisymmetric Exponential Model, we may reduce the problem to an one-dimensional problem in terms of γ. For 4-parameter Gompertz Model, the problem can be reduced to one-dimension on δ. For Extended Chapman–Richards Model, however, this approach may not lead to desired situation. This is because the substitution formula $w(n)=((\delta - y(n))/\alpha)^{1/\gamma}$ involves three parameters and thus the dimension of the problem can only be reduced from four to three. This approach may not work for Reparameterization of Gompertz Model either, because the substitution $w(n) = \ln(y(n) - \delta)$ tends to $-\infty$ when n increases. Hence, we do not recommend this approach for the Extended Chapman–Richards Model and the Reparameterization of Gompertz Model.

4. Preliminary numerical experiments

In this section, we shall report a preliminary numerical experiment using the Exponential model. Our object if to compare three numerical approaches for parameter estimation. These are: Method 1: the classical Newton's method; Method 2: the approach presented in the previous section; and Method 3: the combination of Method 2 and Method 1, which involves using the results of Method 2 as initial guess and excuting Method 1.

The experiment is first based on simulated data generated from

$$y(n) = Z(n, \boldsymbol{\theta}) + x(n), \quad n = 1, 2, \ldots, 200,$$

where

$$Z(n, \boldsymbol{\theta}) = \theta_1 + \theta_2 \exp(-\theta_3 n)$$

with $\theta_1 = 1$, $\theta_2 = 5$, and $\theta_3 = 0.1$. The error terms $x(n)$ is generated using normal distribution $N(0, \sigma)$ with two different values for σ: $\sigma = 0.1 \times \theta_1$ and $\sigma = 0.001 \times \theta_1$. For Method 2 and Method 3, the

Table 3
Numerical results from the simulated data

$\sigma = 0.001 \times \theta_1$	Initial Guess	Estimate	No. of Newton Iter.
Method 1	(1.1, 4.9, 0.09)	(1.0000, 5.0003, 0.1000)	4
Method 1	(1.5, 4.5, 0.2)	—	
Method 2		(0.9996, 5.0020, 0.0999)	
Method 3		(1.0000, 5.0003, 0.1000)	2
$\sigma = 0.1 \times \theta_1$	Initial Guess	Estimate	No. of Newton Iter.
Method 1	(1.1, 4.9, 0.09)	(1.0063, 5.0063, 0.0997)	4
Method 1	(1.5, 4.5, 0.2)	—	
Method 2		(0.9981, 5.3731, 0.1032)	
Method 3		(1.0063, 5.0063, 0.0997)	3

integer k is chosen to be $k = [N/10]$. Thus for this case, $k = 20$. The termination criteria for both Method 1 and Method 3 is

$$\left\| \frac{\partial}{\partial \boldsymbol{\theta}} \sum_{n=1}^{N} [y(n) - Z(n, \boldsymbol{\theta})]^2 \right\| \leqslant 10^{-4}. \tag{31}$$

For Method 1 we used two sets of initial guess — a good initial guess (1.1, 4.9, 0.09) and a poor initial guess (1.5, 4.5, 0.2). It is clear from Table 3 that this method does not work when the initial guess is poor. We also see that when error term is small (when $\sigma = 0.001 \times \theta_1$) Method 2 produces good estimate for parameters. However, when the error term is large (when $\sigma = 0.1 \times \theta_1$) the result of Method 2 alone is not satisfactory. Method 3 clearly has the best performance in either case, supporting our earlier arguments. All numerical results are listed in Table 3, where "—" means the convergence does not occur after 1000 iterations.

We now apply Method 3 to a set of real data presented in Table 4. These are the annual records of 400 m run from the year of 1860 to 1988 listed in Dargahi-Noubary [3]. For this data set Method 2 did not provide a good initial guess. Examination of error term $x(n)$ revealed that this was due to large variation and few records such as those that were set during the World Wars which are not quite consistent with the rest of the data. Thus, Method 3 failed here. We think this was also the reason for failure of attempt made by Smith [8] to predict the altimate record for mile race. To overcome this difficulty we suggest smoothing the data before applying Method 3. This will decrease the error variance and will retain the slowly varying component of the data. Here we applied the simple binomial smoothing given below

$$u(n) = [y(n) + y(n+1)]/2, \quad n = 1, 2, \ldots, N-1 \tag{32}$$

and then applied Method 3 to $\{u(n)\}$. We may also apply second-order smoothing, namely

$$v(n) = [y(n) + 2y(n+1) + y(n+2)]/4, \quad n = 1, 2, \ldots, N-2 \tag{33}$$

and apply Method 3 to $\{v(n)\}$. We refer to (32) and (33) as first-order and second-order binomial smoothing, respectively. In what follows, we present a brief study on the first-order and second-order smoothed data sets, namely $\{u(n)\}$ and $\{v(n)\}$.

Table 4
Data for 400 m run (time is in seconds): 1860 – 1988, from Dargahi-Noubary [3]

Year	Time	Year	Time	Year	Time	Year	Time	Year	Time
1860	53.7	1861	50.2	1862	53.2	1863	51.7	1864	51.7
1865	50.2	1866	52.5	1867	51.4	1868	50.0	1869	51.9
1870	50.7	1871	50.2	1872	49.5	1873	50.3	1874	50.2
1875	50.5	1876	50.5	1877	50.1	1878	51.3	1879	48.9
1880	49.3	1881	48.3	1882	49.9	1883	49.0	1884	48.9
1885	48.5	1886	49.5	1887	49.9	1888	49.7	1889	48.2
1890	48.7	1891	49.1	1892	49.2	1893	48.9	1894	48.7
1895	48.2	1896	48.5	1897	48.7	1898	48.5	1899	49.1
1900	47.5	1901	49.3	1902	49.3	1903	48.7	1904	48.9
1905	48.2	1906	48.5	1907	48.5	1908	47.9	1909	48.3
1910	48.5	1911	48.5	1912	47.7	1913	46.9	1914	48.1
1915	47.7	1916	47.1	1917	48.7	1918	47.3	1919	48.9
1920	48.1	1921	47.7	1922	47.7	1923	47.9	1924	47.4
1925	47.6	1926	48.3	1927	47.5	1928	47.0	1929	47.4
1930	47.6	1931	47.1	1932	46.1	1933	46.6	1934	46.5
1935	46.8	1936	46.1	1937	46.6	1938	46.3	1939	46.0
1940	46.4	1941	46.0	1942	46.6	1943	47.5	1944	47.5
1945	46.7	1946	45.9	1947	45.9	1948	45.7	1949	46.2
1950	45.8	1951	46.0	1952	45.9	1953	45.9	1954	46.1
1955	45.4	1956	45.2	1957	46.0	1958	45.4	1959	45.8
1960	44.9	1961	45.7	1962	45.5	1963	44.6	1964	44.9
1965	45.5	1966	44.7	1967	44.5	1968	43.8	1969	44.4
1970	44.9	1971	44.2	1972	45.0	1973	45.2	1974	45.2
1975	44.93	1976	44.26	1977	45.36	1978	45.47	1979	44.00
1980	44.60	1981	45.12	1982	45.00	1983	45.44	1984	44.27
1985	44.96	1986	44.45	1987	44.32	1988	43.29		

First, assuming that $\{y(n)\}$ has an exponential model, we have the difference equation (3),

$$y(n-1)-y(n)=r(y(n-2)-y(n-1)) \tag{34}$$

which implies that

$$y(n)-y(n+1)=r(y(n-1)-y(n)) \tag{35}$$

From (34) and (35) we see

$$\frac{y(n-1)+y(n)}{2}-\frac{y(n)+y(n+1)}{2}=r\left(\frac{y(n-2)+y(n-1)}{2}-\frac{y(n-1)+y(n)}{2}\right),$$

that is,

$$u(n-1)-u(n)=r(u(n-2)-u(n-1)). \tag{36}$$

This shows that $\{u(n)\}$ also an exponential model with the same constant r. The same is true for the second-order smoothed data $\{v(n)\}$. We can also derive these directly from the formula

$$y(n)=\theta_1+\theta_2\exp(-\theta_3 n)+x(n), \quad n=1,2,\ldots,N$$

which implies that

$$u(n) = \theta_1 + \frac{\theta_2(1+\exp(-\theta_3))}{2}\exp(-\theta_3 n) + \frac{x(n)+x(n+1)}{2}, \quad n = 1,2,\ldots,N-1 \tag{37}$$

and

$$v(n) = \theta_1 + \frac{\theta_2(1+2\exp(-\theta_3)+\exp(-2\theta_3))}{4}\exp(-\theta_3 n) + \frac{x(n)+2x(n+1)+x(n+2)}{4}, \quad n = 1,2,\ldots,N-2. \tag{38}$$

Eqs. (37) and (38) show that $u(n)$ and $v(n)$ have exponential models and that the values of θ_1 and θ_3 are the same for $y(n)$, $u(n)$, and $v(n)$. The values of the second parameter are different. If θ_2' and θ_2'' denote the second parameter for $u(n)$ and $v(n)$, respectively, then

$$\theta_2 = 2\theta_2'/(1+\exp(-\theta_3)) \tag{39}$$

or

$$\theta_2 = 4\theta_2''/(1+2\exp(-\theta_3)+\exp(-2\theta_3)). \tag{40}$$

The use of the smoothing procedure in general can also be viewed from a different angle. For example, suppose the data set $\{y(n)\}$ satisfies the difference equation

$$y(n) - y(n+2) = r(y(n-1) - y(n+1)), \quad \forall n, \tag{41}$$

where $r \in (0,1)$ is a constant. Then clearly the subsitution $u(n) = [y(n) + y(n+1)]/2$ satisfies the difference equation

$$u(n) - u(n+1) = r(u(n-1) - u(n)), \quad \forall n$$

and thus has an exponential model. Similarly, if $\{y(n)\}$ satisfies the difference equation

$$\frac{y(n-1)+y(n)}{2} - \frac{y(n+1)+y(n+2)}{2} = r\left[\frac{y(n-2)+y(n-1)}{2} - \frac{y(n)+y(n+1)}{2}\right], \quad \forall n \tag{42}$$

then the second-order smoothed data $\{v(n)\}$ has an exponential model. Note that difference equation (41) is based on the ratio of improvements between terms that are two years apart, while (42) is based on the ratio of improvements between averages of two pairs.

Returning to our numerical example based on the data listed in Table 4, we tried the following four procedures:

(I) Applied Method 3 to original data $\{y(n)\}$. This method failed as reported previously in this section.

(II) Applied Method 3 to the first-order smoothed data $\{u(n)\}$ and then use the formula (39) to convert θ_2' back to θ_2. This procedure worked out successfully. The estimate of parameters is listed in Table 5.

Table 5
Numerical results from the records of 400m run

Procedure	Data size	k	Estimate
(I)	129	12	—
(II)	128	12	(37.6219, 13.9014, 0.0056)
(III)	127	12	(37.6224, 13.8839, 0.0056)
(II)–(IV)	(II)128–(IV)129		(38.1673, 13.4402, 0.0060)
(III)–(IV)	(III)127–(IV)129		(38.1673, 13.4402, 0.0060)

(III) Applied Method 3 to the second order smoothed data $\{v(n)\}$ and then use formula (40) to convert θ_2'' back to θ_2. Again, it successfully obtained the estimate which is also listed in Table 5.

(IV) We then applied the Newton's method directly to the original data $\{y(n)\}$ using the results of procedures (II) and (III) as initial guesses. The estimates obtained are also listed in Table 5, where "(II)–(IV)" means "procedure (IV) using results of procedure (II) as initial guess" and "(III)–(IV)" has a similar meaning.

From Table 5 we see that the results of procedures (II)–(IV) are all very close, supporting the theoretical justification for using the smoothing procedures. The estimates listed in Table 5 also suggest that the lower limit for the time to run 400 m is about 38 s.

It is interesting to note that Chatterjee and Chatterjee [2] have analyzed the winning time in the men's 400 m run in the Olympic Games from 1900 to 1976. They have fitted the exponential model using nonlinear model fitting and Jack-Knife procedure. Their analysis have predicted an estimate for the lower bound (θ_1) equal to 38.975. Since record for 400-meter run was set in the year 1988 and is included in our analysis, estimates given in Table 5 seem reasonable.

Using the estimate obtained from procedure (IV) as listed in Table 5, we are able to give a prediction on the winning time in 400 m run for the years of 2050 and 2100. These predictions, computed using the exponential model are, respectively, 42.44 s (for 2050) and 41.33 s (for 2100), suggesting that in long run the winning times will approach a lower limit.

Finally, we should like to mention that the Method 2, when used as an approach to provide initial estimation for parameters, has some advantages over existing methods such as the one offered in Ratkowski [6]. His approach [6, p. 160], when applied to an exponential model, can be described in steps as follows:

Method 2' (Ratkowski, [6]).

Step 1:

1.1. Select a step size ε, $0 < \varepsilon < 1$, e.g. $\varepsilon = 0.05$. Let $\theta_3^{(0)} = \varepsilon$.

1.2. Apply the linear least-squares approach to find $\theta_1^{(0)}$ and $\theta_2^{(0)}$ by minimizing

$$\sum_{n=1}^{N} \left[y(n) - \theta_1 - \theta_2 \exp\left(-\theta_3^{(0)} n\right) \right]^2 .$$

Set $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \theta_3^{(0)})$.

Table 6
Comparision of Method 2 with Method 2′

Method	ε	Iter. in Method 2′	Initial Guess	No. of Newton Iter.
Method 2			(0.9981, 5.3731, 0.1032)	3
Method 2′	0.003	33	(1.0013, 4.9133, 0.0960)	3
Method 2′	0.03	4	(0.9926, 4.7599, 0.0900)	4
Method 2′	0.04	3	(0.9764, 4.4756, 0.0800)	4
Method 2′	0.05	3	(1.0067, 5.0142, 0.1000)	2
Method 2′	0.06	3	(1.0304, 5.4765, 0.1200)	5
Method 2′	0.14	2	(1.0494, 5.8813, 0.1400)	7
Method 2′	0.15	2	(1.0577, 6.0664, 0.1500)	—

1.3. Compute the weighted least-squares error as

$$s^{(0)} = \sum_{n=1}^{N} \left[\frac{Z\left(n, \boldsymbol{\theta}^{(0)}\right)}{y(n)} - 1 \right]^2 .$$

Step 2: For $k = 1, 2, \ldots$;
2.1. Let $\theta_3^{(k)} = \theta_3^{(k-1)} + \varepsilon$.
2.2. Apply the linear least-squares approach to find new values $\theta_1^{(k)}$ and $\theta_2^{(k)}$ by minimizing

$$\sum_{n=1}^{N} \left[y(n) - \theta_1 - \theta_2 \exp\left(-\theta_3^{(k)} n\right) \right]^2 .$$

Set $\boldsymbol{\theta}^{(k)} = (\theta_1^{(k)}, \theta_2^{(k)}, \theta_3^{(k)})$.
2.3. Compute the weighted least-squares error as

$$s^{(k)} = \sum_{n=1}^{N} \left[\frac{Z\left(n, \boldsymbol{\theta}^{(k)}\right)}{y(n)} - 1 \right]^2 .$$

2.4. If $s^{(k)} > s^{(k-1)}$ then take $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k-1)}$ as the initial value for the Newton's method and stop, otherwise continue with the next iteration.

We compared Method 2 with Method 2′ as approaches to provide initial values for Newton's method. The major advantage of Method 2 over Method 2′ is that Method 2 does not require the user to provide an step size ε to start with. In fact, the performance of Method 2′ largely depends on the choice of the step ε. With the simulated data with $\theta_1 = 1$, $\theta_2 = 5$, and $\theta_3 = 0.1$, Method 2′ performed well with the choice $\varepsilon = 0.05$. In fact, it worked better for $\varepsilon = 0.05$ than for $\varepsilon = 0.04$ or 0.03 because in this case the true value $\theta_3 = 0.1$ is a multiple of 0.05. Method 2′ may still work out even if the value of ε gets bigger than 0.1. For example when $\varepsilon = 0.14$, Method 2′ would just accept $\boldsymbol{\theta}^{(0)}$ as the initial guess and the Newton's method does converge with this initial guess. However, if a user happens to choose an ε bigger than 0.15 then Method 2′ fails to provide a working initial guess for the Newton's method. The same happens with the data of 400 m run records, where Method 2′ works out when $\varepsilon \leqslant 0.013$ but fails when $\varepsilon > 0.013$. Therefore, the performance of Method 2′ is

quite sensitive to the choice of ε. In general, a smaller ε often leads to a better initial guess for the Newton's method, but meanwhile means that more iterations are required within the procedure of Method 2′. With Method 2 the user does not have to make such a choice. We also notice that the amount of computation in each iteration of Method 2′ is about the same as the total amount required by Method 2. As an example, in Table 6 we list the numerical results comparing Method 2 with Method 2′ on the simulated data when $\sigma = 0.1 \times \theta_1$. The results on other data sets have demonstrated a similar pattern and therefore are omitted.

5. Concluding remarks

In this paper we established a relationship between a class of difference equations and set of models that are used frequently for prediction of future records. We found that these models are essentially of the same type and can be written in one form or another for a further analysis. In particular, they can all be converted to the exponential model via appropriate substitutions. One very interesting observation worth mentioning relates to the striking resemblance of the above models with those used in population growth. A clear example is the Logistic equation which is sufficiently general and flexible for describing population growth, and any other problem being similar in nature. In fact, this resemblance can be explained in fashion which we think is quite convincing. To clarify this consider a population with a rate of increase greater than zero. It is reasonable to assume that as population increase so does the chance of producing an exceptional athlete capable of breaking record. Note that even when population is not physically increasing, it is possible to have an increase in population of athletes. So the problem of modeling and prediction of records may be looked upon as a population growth problem. Thus any model suitable for describing the population growth may also be used to describe records. Note that one advantage of this connection is that many powerful models have already been introduced and used by scientists for pupolation growth. In fact, a large number of established models of this type are based on difference equations involving improvements or some functions of them. We also noticed that all the models discussed in this paper can be written in an equivalent form as

$$y(n) = Z(n, \theta) + x(n)$$

where $Z(n, \theta)$ is the deterministic part being of main concern for prediction. Here $x(n)$ is only needed for evaluation of the accuracy of the prediction. Note that the usual assumption of i.i.d. for $x(n)$'s is not a realistic one. In fact, as is demonstrated in Dargahi-Noubary [3] the records are highly correlated and this makes the analysis more complicated and less reliable. So rather than making a convenient assumption regarding $x(n)$, we decided to focus on models for $Z(n, \theta)$. Further, to avoid failure of the estimating procedure we decided to smooth the data. This, as is well known will reduce the variation and will retain the slowly varying part of data which is more relevant and carries information regarding $Z(n, \theta)$.

Acknowledgements

We like to thank the referees for their valuable comments and suggestions on the earlier version of this paper.

References

[1] D.C. Blest, Focus on sport lower bounds for athletic performance, The Statistician 45 (2) (1996) 243–253.
[2] S. Chatterjee, S. Chatterjee, New lamps for old: An exploratory analysis of running times in Olympic games, Applied Statistics 31 (9) (1982) 14–22.
[3] G.R. Dargahi–Noubary, An envelope function model for forecasting atheletic records, J. Forecasting 13 (1994) 11–20.
[4] L. De Haan, Estimation of the minimum of a function using order statistics, J. Amer. Statist. Assoc 76 (1981) pp 467–469.
[5] J.J. McKeown, D. Sprevak, Parameter estimation versus curve fitting: new lamps for old, The Statistician 41 (1992) 357–361.
[6] D.A. Ratkowsky, Nonlinear regression modeling: a unified practical approach, Dekker, New York, 1983.
[7] D.A. Ratkowsky, Hankbook of Nonlinear Regression Models, Dekker, New York, 1990.
[8] L.R. Smith, Forecasting records by maximum likelihood, J. Amer. Statist Assoc 83 (1988) 331–338.
[9] P. Tryfos, R. Blackmore, Forecasting records, J. Amer. Stat. Assoc 80 (1985) 46–50.
[10] R. Wootton, J.R. Royston, New lamps for old, Appl. Statist. 32 (1983) 88–89.